Fundamentos de IA y Machine Learning

Lección 1: Introducción al Aprendizaje Automático

eip ESCUELA INTERNACIONAL POSGRADOS

ÍNDICE

Introducción al Aprendizaje Automático

PRESENTACIÓN Y OBJETIVOS

El Aprendizaje Automático engloba una gran cantidad de conceptos y procedimientos que son difíciles de entender sin una perspectiva introductoria a esta rama de la ciencia. En esta lección se pretende abordar estos conceptos introductorios para el posterior análisis de los datos y una primera toma de contacto con la visualización de estos.

Objetivos

Al finalizar esta lección serás capaz de:

- Saber que es la ciencia de datos y el aprendizaje automático.
- Entender la diferencia entre los distintos problemas existentes.
- Comprender las distintas etapas en las que se divide un problema de *Machine Learning*.
- Introducir los primeros conceptos sobre el preprocesamiento de los datos.
- Visualizar los datos de forma gráfica.

1. Introducción

1.1. Motivación

Actualmente, vivimos rodeados de un mundo que gira en torno a datos cuya cantidad crece de forma exponencial con el paso del tiempo.

Podemos imaginar la cantidad de datos que se generan en ramas como:

- **Ciencia**: astronomía, genómica, medioambiente, hidrología, etc.
- **Industria y energía**: redes de sensores, Internet de las cosas, gestión de parques eólicos, previsión de demanda, ciudades inteligentes, etc.
- **Ciencias sociales y humanidades**: libros digitalizados, documentos históricos, datos sociales, etc.
- **Entretenimiento**: sistemas de recomendación, contenidos digitales, búsquedas multimedia, etc.
- **Medicina**: examen de imágenes médicas, sistemas expertos, etc.
- **Financias y negocios**: transacciones de mercados automatizadas, etc.

Hasta hace poco, era impensable procesar la cantidad de datos debido a distintas restricciones. Sin embargo, hoy en día, la mejora de las tecnologías de las bases de datos, la reducción del coste del *hardware* de almacenamiento, y el aumento del ancho de banda, la capacidad de procesado y *software* científico nos brinda la capacidad de pasar de información a conocimiento.

1.2. Ciencia de datos

En este contexto, surge lo que conocemos como ciencia de datos. Ésta se define como el ámbito de conocimiento que engloba las habilidades asociadas al procesamiento de datos.

José Antonio Guerrero, uno de los mejores científicos de datos del mundo, define lo que sería un científico de datos:

"Es una persona con fundamentos en matemáticas, estadística y métodos de optimización, con conocimientos en lenguajes de programación y que además tiene una experiencia práctica en el análisis de datos reales y la elaboración de modelos predictivos. De las tres características quizás la más difícil es la tercera; no en vano la modelización de los datos se ha definido en ocasiones como un arte. Aquí no hay reglas de oro, y cada conjunto de datos es un lienzo en blanco."

2. APRENDIZAJE AUTOMÁTICO

El aprendizaje automático o aprendizaje de máquina (*machine learning* en inglés) se define como el campo de estudio que proporciona a los ordenadores la capacidad de aprender sin haber sido explícitamente programados para ello. Equivale a "aprender de los datos" con el fin de extraer el conocimiento necesario según diferentes propósitos.

Aprender de los datos implica que el aprendizaje automático se sitúe entre diferentes ramas que pertenecen a la inteligencia artificial, la estadística y las matemáticas.

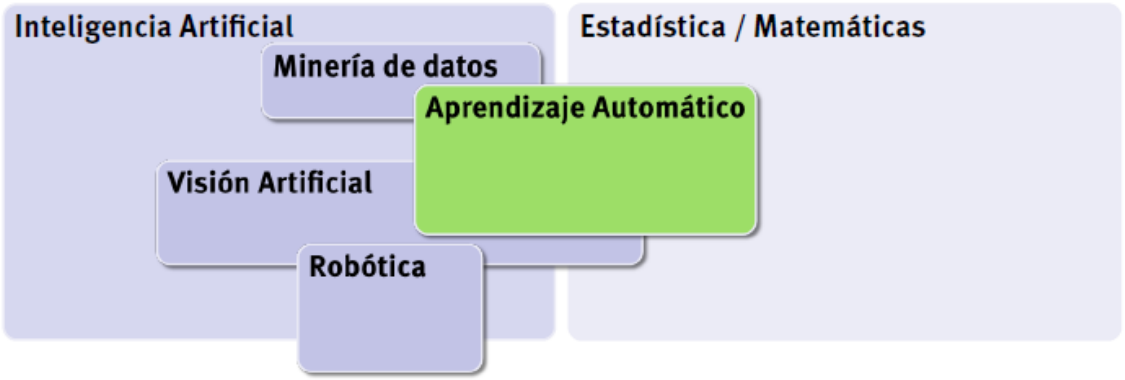


Figura 1.1: Aprendizaje automático.

2.1. Presentación de los datos

La información con la que trabajamos se puede presentar de distintas formas: tablas, imágenes, audios, vídeos, mapas en dos dimensiones, espacios tridimensionales... En definitiva, cualquier forma de organizar los datos de forma ordenada.

Cada uno de esos formatos tiene unas características especiales, que hacen que los algoritmos diseñados para ellos sean propios y específicos. Sin embargo, hay forma de transformar los datos de un formato a otro. Así, por ejemplo, en el procesado de audios, estos se suelen transformar en espectrogramas de dos dimensiones para poder ser tratados como imágenes.

Lo más común es la utilización de datos organizados en tablas o matrices, de forma que se pueden utilizar notaciones vectoriales para referirnos a los datos. De esta forma:

- **Cada fila de la tabla es un patrón, instancia u observación.**
- **Cada columna es una característica, atributo o variable de entrada del patrón observado.**
- **En general, habrá una columna objetivo o variable de salida, que será el valor que se pretende predecir, aunque esto no siempre ocurre como veremos más adelante.**

A continuación, se expone un ejemplo totalmente ficticio de base de datos que recopila información de pacientes a partir de cuatro atributos (Edad, Fuma, Deporte, Comida saludable) para determinar si tiene riesgo de padecer una enfermedad.

Edad	Fuma	Deporte	Comida saludable	Enfermedad
28	No	Sí	Sí	**No**
16	No	No	No	**No**
45	Sí	Sí	No	**Sí**
65	Sí	No	No	**Sí**

2.2. Clasificación de los métodos en aprendizaje automático

Dependiendo de la variable objetivo definida en la sección anterior, los métodos de aprendizaje automático se pueden dividir en a) aprendizaje supervisado, b) no supervisado o c) semisupervisado.

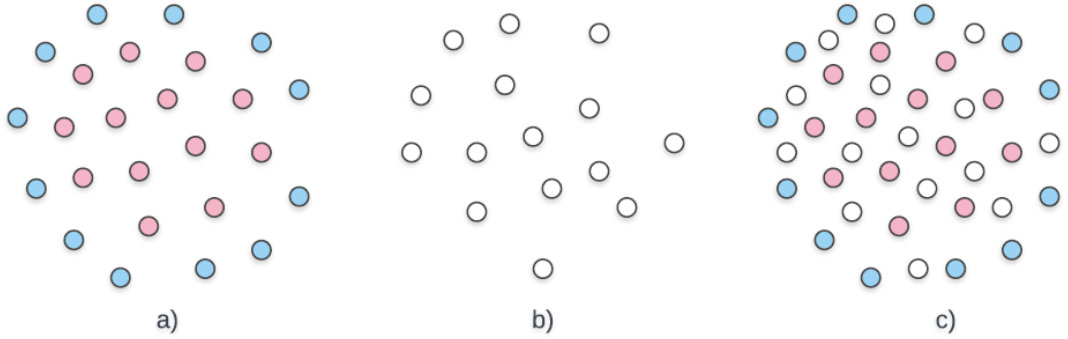
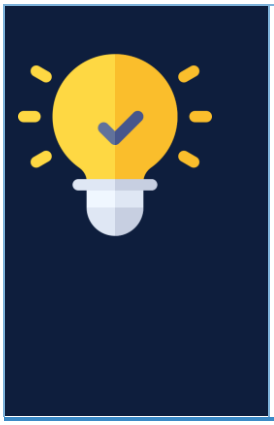


Figura 2.1: Métodos aprendizaje automático

2.2.1. Aprendizaje supervisado

Es el tipo de aprendizaje más común. Un problema puede ser resuelto por un algoritmo de aprendizaje supervisado cuando cada patrón u observación (denotado como un vector de características $\boldsymbol{X}$) está etiquetado por una clase o por un valor Y, es decir, tiene asociada una variable objetivo como definimos en la sección anterior.

De esta forma, el principal objetivo es diseñar un algoritmo automático que aprenda un conjunto de reglas a partir de una parte del conjunto de datos, denominada de entrenamiento (*train set* en inglés), con el objetivo de aproximar la salida real en el conjunto de generalización (*test set)*.

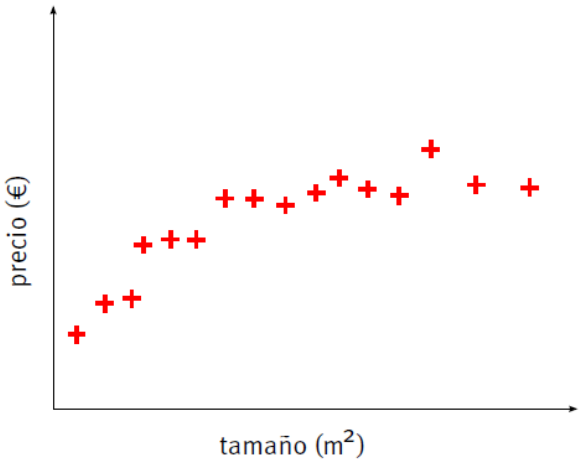


Importante

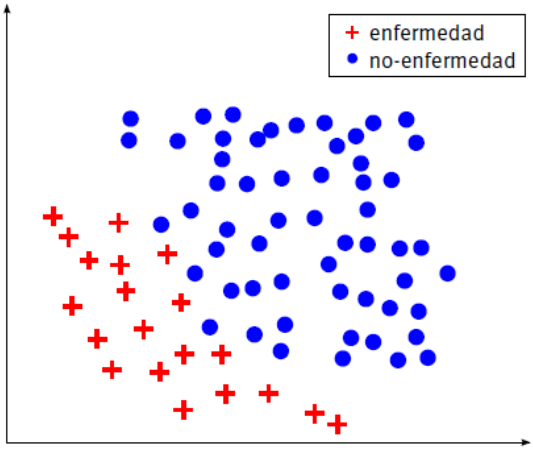
La lección 2 de esta asignatura tratará con más detalle el concepto de conjunto de entrenamiento y generalización, y el porqué de su uso.

En este grupo, se realiza la siguiente división:

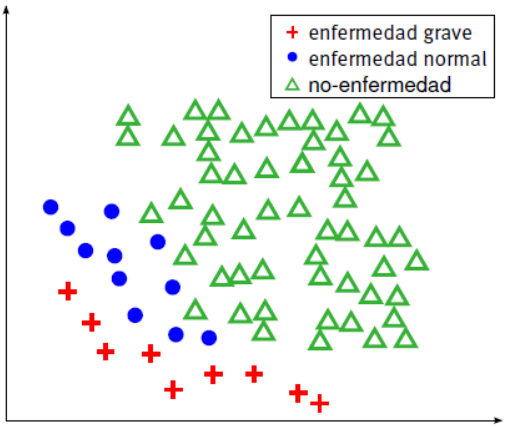
- **Regresión**: la variable de salida Y es un valor real. Un ejemplo de algoritmo en este grupo sería la regresión lineal.



- **Clasificación**: la variable de salida Y es un valor discreto, nominal o categórico. Los métodos de clasificación son los más extendidos en la literatura y algunos de los que se verán en la Lección 3 y 4 son: regresión logística, vecinos más cercanos, árboles decisión, maquinas de vectores soporte, redes neuronales artificiales y metaclasificadores o *ensembles*.



Clasificación o regresión ordinal: la variable de salida Y es categórica y además existe un orden entre las clases. Como se puede observar, se trata de una mezcla entre la clasificación y la regresión. El problema es que este orden no esta cuantificado, es decir, no se puede determinar la distancia entre las clases. Así por ejemplo, si tenemos un clasificador que intenta categorizar un conjunto de patrones en las clases *niño, adolescente, adulto* o *anciano*, se sabe que clasificar un *niño* como un *anciano* debería ser más penalizado que clasificarlo como *adolescente*. Es decir, podemos determinar que la distancia entre *niño* y *adolescente* es menor que la distancia entre *niño* y *anciano*, pero no se puede determinar si la distancia entre *niño* y *adulto*, es el doble de la distancia entre *niño* y *adolescente*.



2.2.2. Aprendizaje no supervisado

En este caso, por cada patrón de características de entrada $\boldsymbol{X}$, no se conoce una variable objetivo. La idea subyace en encontrar distintos grupos de patrones que presenten una estructura similar, determinar la distribución de los datos, o bien, proyectar los datos en un espacio de menor dimensionalidad para poder visualizarlos. Este paradigma incluye a los algoritmos de agrupamiento o *clustering* cuyo objetivo es encontrar grupos de patrones teniendo en cuenta sus características de entrada. Los algoritmos de *clustering* se dividen en particionales, jerárquicos y basados en densidad.

Importante

La lección 5 de esta asignatura tratará con más detalle el concepto de *clustering* así como los distintos tipos de algoritmos dentro de este paradigma.

2.2.3. Aprendizaje semi-supervisado

Este aprendizaje tiene lugar cuando el conjunto de datos presenta un subconjunto que está etiquetado, y un gran subconjunto de datos no etiquetado. Esto se produce cuando es difícil etiquetar la totalidad de los datos, ya sea porque el proceso tiene un gran coste o porque depende de un experto, del que no se dispone en determinadas ocasiones. Estos algoritmos tratan de explorar la estructura de los datos sin etiquetar para la generación de modelos predictivos que funcionen mejor que aquellos que sólo utilizan el conjunto etiquetado.

2.3. Fases del proceso de construcción de un modelo de Aprendizaje Automático

Se pueden distinguir las siguientes etapas en la construcción de un modelo de aprendizaje automático:

1. Integración y recopilación: compresión del dominio de aplicación del problema, identificación de conocimiento a priori y creación de un almacén de datos. (Lección 1 de la asignatura).
2. Preprocesamiento: selección de datos, limpieza, reducción y transformación. Las etapas 1 y 2 serían las que conllevan un mayor esfuerzo. (Lección 1).

3. Selección de la técnica, aplicación de la misma y entrenamiento. (Lección 2, 3, 4 y 5).
4. Evaluación, interpretación y presentación de los resultados obtenidos. (Lección 2, 3, 4 y 5).
5. Difusión y utilización del nuevo conocimiento. (Lección 2, 3, 4 y 5).

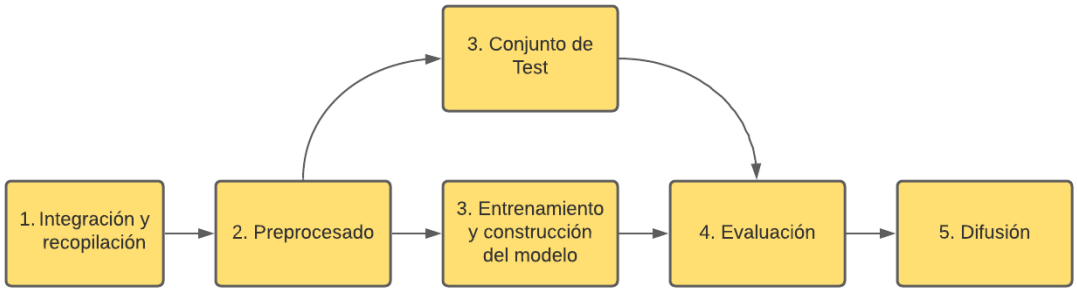


Figura 2.2 Fases del Aprendizaje Automático

3. Preprocesado de los datos

A menudo, los resultados van a depender más de la calidad de los datos con relación al problema, que de la parte de la generación del modelo.

Es frecuente hablar de ruido en los datos o de la relevancia de las variables, pero no se debe olvidar que esto es siempre respecto a un objetivo. Es decir, una variable de entrada puede ser ruido en un problema, mientras que puede ser información muy útil y relevante en otro.

Existen distintas tareas de preprocesado.

3.1. Selección de variables o características

Como se ha comentado previamente, cada patrón viene determinado por un conjunto de características o variables.

Para obtener una buena base de datos es necesario aplicar:

- **Extracción de características:** determinar que variables necesitamos para la caracterización de un problema. Por ejemplo, al procesar datos multimedia se extraen características que permiten construir vectores de tamaño fijo necesarios para los modelos.

- **Selección de características:** este proceso consiste en descartar aquellas características que no son necesarias en nuestro problema. Por ejemplo, variables con un valor constante, variables que carecen de significado en un problema (color de ojos para determinar si una persona tiene riesgo de sufrir un ataque cardiaco), etc.

3.2. Limpieza de datos

Se realizan operaciones sobre los datos, ya sean en variables o patrones. Por ejemplo:

- **Recuperación de valores perdidos**: puede ocurrir que en algunos casos no tengamos un valor para una variable en un patrón. Existen distintas técnicas de imputación de datos. Una muy común es utilizar la media de los valores de la variable para imputarlos. Otra, más sofisticada sería utilizar la media de los valores de los patrones de la clase de ese patrón.

- **Tratamiento de valores anómalos** (*outliers)* **o inconsistencias**: en muchas ocasiones se pueden presentar valores que pueden ser casos "extraños" o bien que han sido mal anotados. Por ejemplo, supongamos una variable que mide la altura de las personas en centímetros y nos encontramos con un valor de 1800. Posiblemente, el error está en que al obtener el dato se ha anotado un "0" por lo que el valor real sería de 180. Esto es sólo un ejemplo y el tratamiento de valores anómalos depende del problema a tratar y de la variable en cuestión.

- **Tratamiento del ruido:** dependiendo del problema nos puede interesar suavizar ruido o aumentar el ruido de las variables.

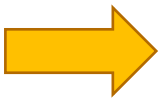
3.3. Transformación de datos

Esta tarea es esencial para el correcto desarrollo del método de aprendizaje automático. Cómo se ha comentado anteriormente, cada patrón o instancia está compuesta por un conjunto de variables. Estas variables pueden ser numéricas, enteras o reales, lógicas, categóricas, fechas, etc.

Algunos de preprocesados más importantes son:

- **Binarización de variables**: puesto que muchos algoritmos de *machine learning* trabajan con valores numéricos, los datos categóricos deben ser transformados. El método más usado es el *One-Hot-Encoding*, y se utiliza cuando no existe una relación entre las categorías. En este método, una variable se convierte en tantas como valores distintos existen. Pro ejemplo, la categoría "animal" que tiene como posibles valores "gato", "perro" y "conejo" quedaría de la siguiente forma.

Animal
Perro
Gato
Conejo
Perro



Perro	Gato	Conejo
1	0	0
0	1	0
0	0	1
1	0	0

- **Escalado de variables**: en la mayoría de ocasiones el rango de las variables debe ser el mismo para poder generar modelos con significado. Por ello, es necesario realizar el escalado y los más conocidos son:

  - **Estandarización N(0,1)**: normaliza los valores de una variable a una distribución con media 0 y varianza igual a 1. Para ello, se calcula la media y la desviación de una característica j, y para cada uno de los patrones $\boldsymbol{x}_i$, se aplica la siguiente transformación:

$$x'_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}$$

  - **Escalado [0, 1]:** se escalan los valores en el intervalo [0, 1].

$$x'_{i,j} = \frac{x_{i,j} - min(x_j)}{max(x_j) - min(x_j)}$$

- **Transformar fechas:** las variables de tipo fecha suelen ser cadenas de texto que los algoritmos de aprendizaje automático no son capaces de comprender. Una solución es restar la fecha actual a la fecha de nuestra variable para obtener un número de años que ya podría ser tratado por estos algoritmos.

3.4. Reducción de la dimensionalidad

Es distinto a la selección de características. Se basa en emplear técnicas como el análisis de componentes principales (PCA) para obtener combinaciones lineales de variables que reduzcan la dimensión de los datos.

3.5. Tratamiento del desbalanceo

Puede ocurrir que, en problemas supervisados de clasificación, exista un número de patrones distinto por cada clase. Esto da lugar al desbalanceo, es decir, que el número de patrones de una clase sea muy grande con respecto a otra. En este sentido, los algoritmos de aprendizaje automático tienden a "aprender" a clasificar bien las clases mayoritarias. Para evitar este problema tenemos dos alternativas:

- **Técnicas de *oversampling***: se generan patrones sintéticos en aquellas clases con menor número de patrones.
- **Técnicas de *undersampling***: se reduce el número de patrones de las clases mayoritarias.

4. VISUALIZACIÓN DE LOS DATOS

Los distintos algoritmos de aprendizaje automático buscan estadísticamente las relaciones existentes entre los patrones para las distintas variables. Si todos los patrones tuvieran los mismos valores, el conjunto tendría información muy reducida. Por otro lado, si no existiese ninguna relación entre sus variables, tampoco sería posible la extracción de información útil.

Un científico de datos tiene la tarea de determinar cómo se distribuyen los datos en las distintas variables, y qué relación tienen las variables entre ellas. Precisamente, esto es lo que los algoritmos realizan de forma más compleja y abstracta. Esta visualización se puede llevar a cabo mediante histogramas, gráficos de dispersión (*scatter-plot)* y mapa de calor (*heatmap)*.

4.1. Distribución de los datos

Una forma de visualizar la distribución de los datos es mediante el uso de histogramas. Estos consisten en la representación gráfica en forma de barras, donde el eje de abscisas consiste en el rango en el cual se encuentran los datos. Este debe ser dividido en distintos intervalos. Mientras que el eje de ordenadas representa la frecuencia de valores en cada uno de esos intervalos.

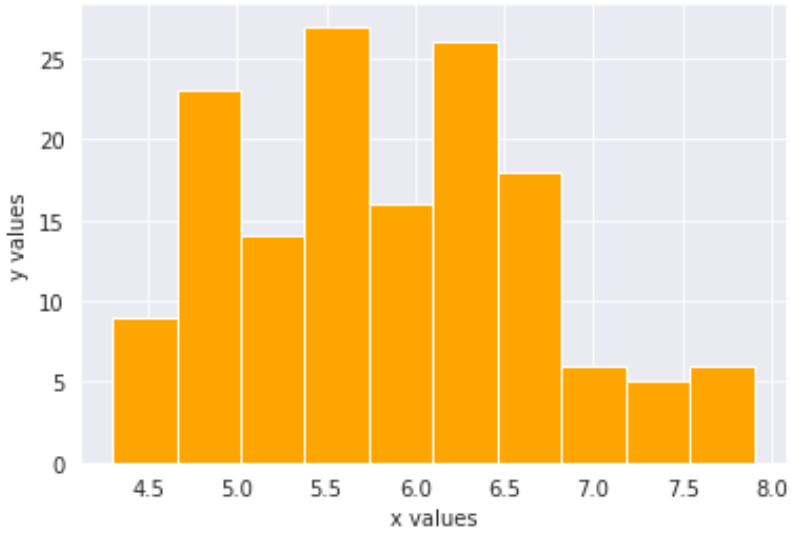


Figura 4.1. Histograma visualización distribución de datos.

Cada variable tiene un valor medio (μ), una desviación típica (σ) y distintos cuartiles (Q1, Q2, Q3) que indican el valor que deja por detrás el 25%, 50% y 75% de los datos. Al cuartil Q2 también se le denomina mediana. Una forma de visualizar esta información gráficamente es haciendo uso de los *boxplots* o diagrama de cajas.

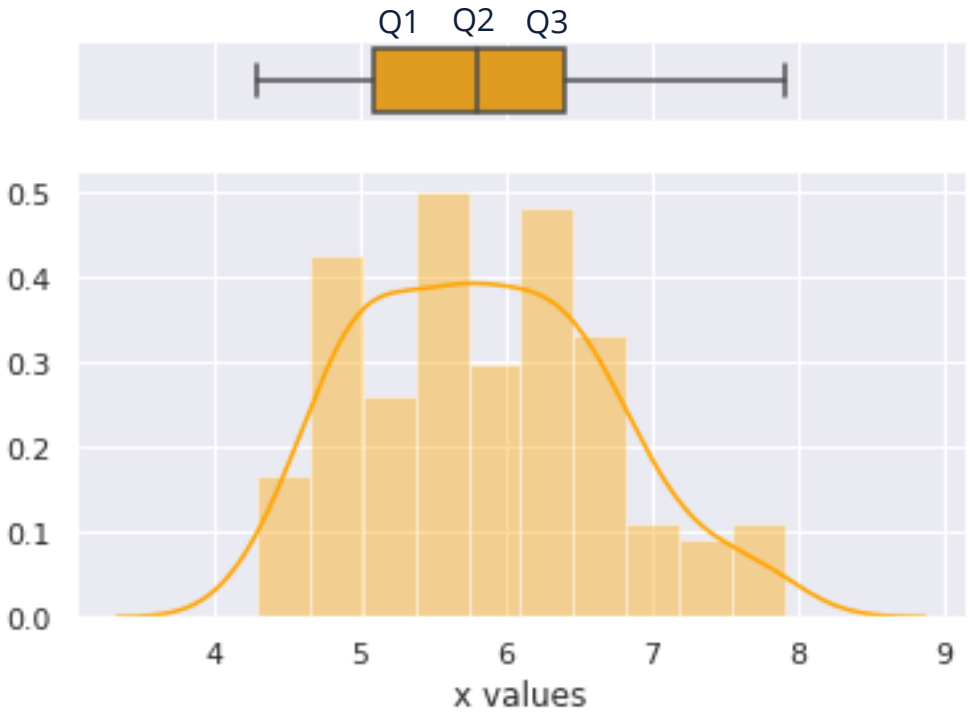


Figura 4.2. Boxplots o diagrama de cajas.

4.2. Correlación entre variables

Con el gráfico de dispersión o *scatter-plot* podemos ver de una manera muy intuitiva como se relacionan dos variables entre sí. A continuación, se muestra un ejemplo de relación entre pares de variables de una base de datos denominada Iris.

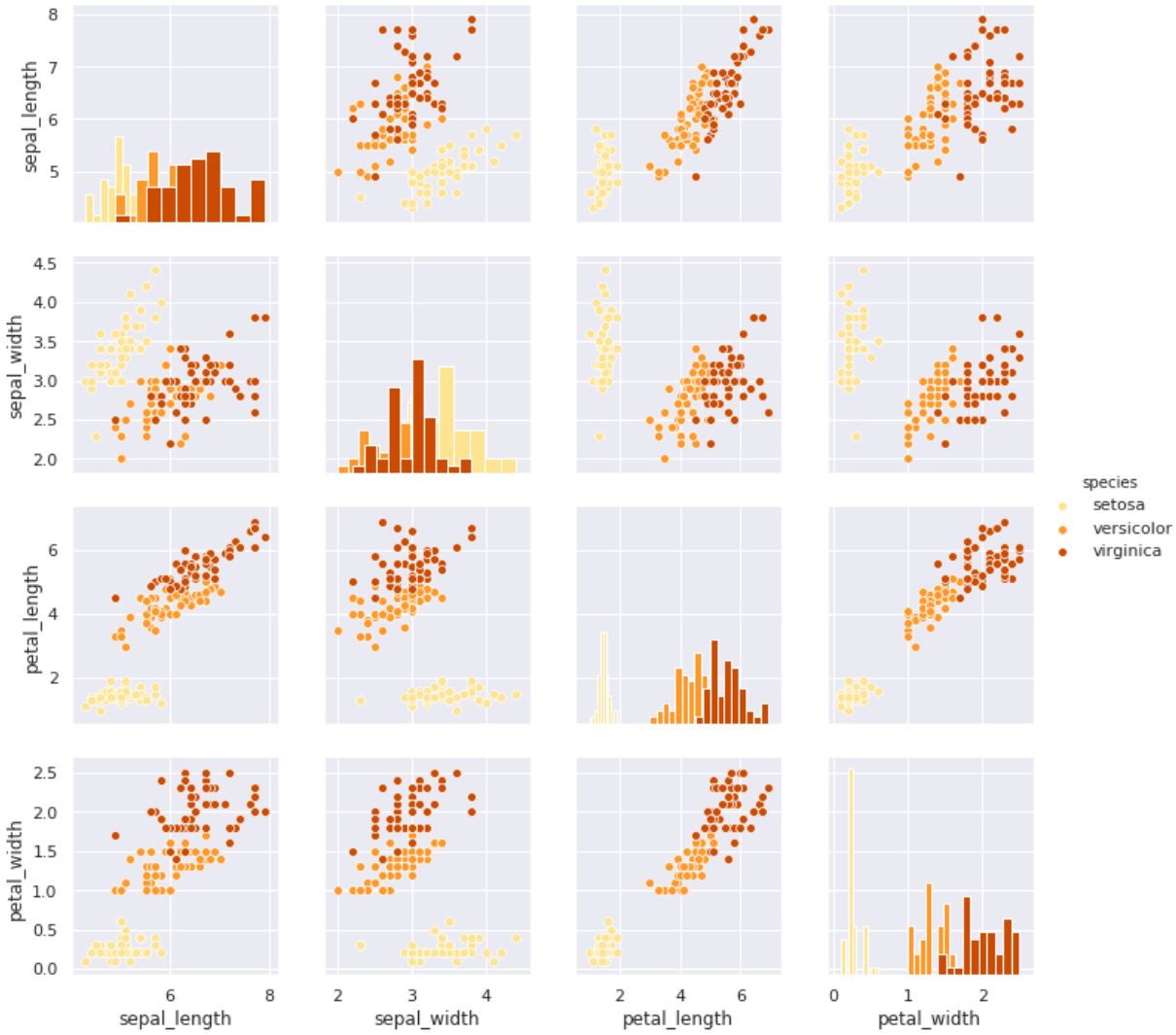
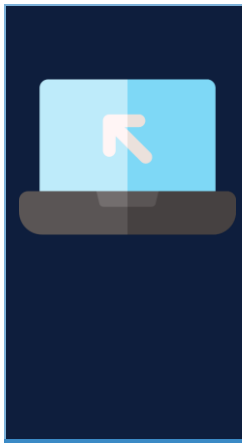


Figura 4.3. Gráfico de dispersión o *scatter-plot*



Para más información

Si visitas el siguiente enlace conocerás en profundidad las características de la base de datos Iris.

https://archive.ics.uci.edu/ml/datasets/iris

Otra forma de mirar la relación entre las variables, es mediante un mapa de calor o *heatmap*. La idea sería mirar la correlación lineal de Pearson entre cada par de variables de nuestro conjunto de datos. Con la base de datos anterior, podemos ver el siguiente mapa de calor.

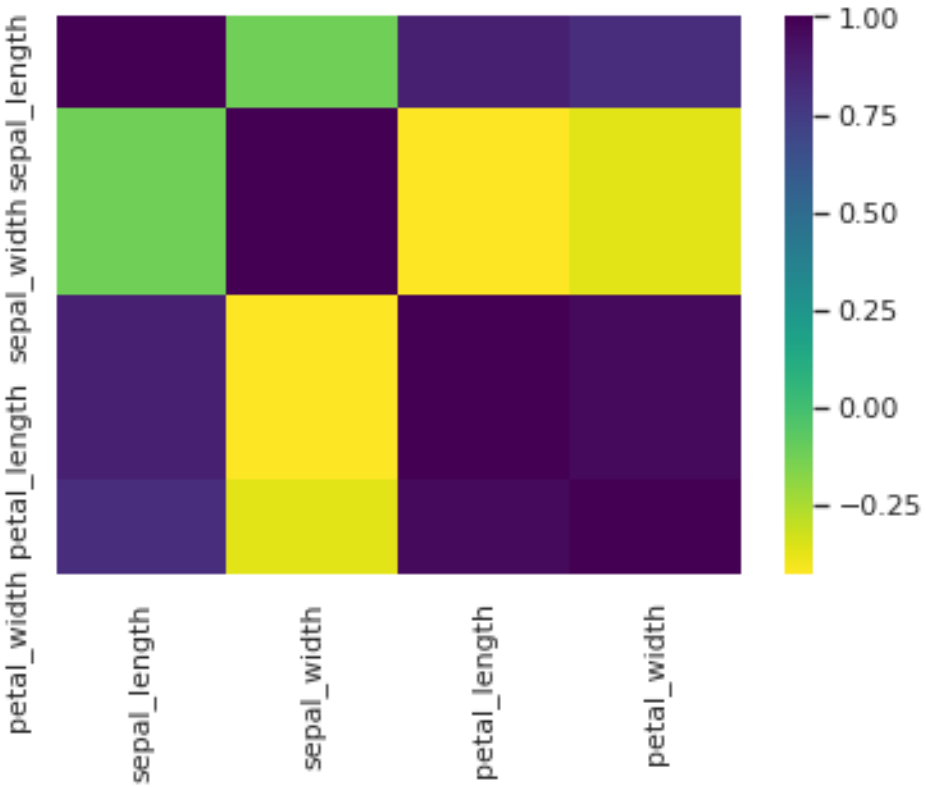


Figura 4.4. Mapa de calor o heatmap

5. PUNTOS CLAVE

Una vez que hemos desarrollado los puntos más importantes de la introducción al aprendizaje automático y en concreto, la fase de preprocesado y visualización, estaremos capacitados para:

- Diferenciar los paradigmas del aprendizaje automático.
- Conocer los tipos de problemas de aprendizaje supervisado.
- Limpiar y cribar la base de datos para posteriores tareas.
- Visualizar los datos de forma intuitiva.